NPS55-78-032

NAVAL POSTGRADUATE SCHOOL

Monterey, California

EXPERIMENTATION MANUAL

PART I: EXPERIMENTATION METHODOLOGY

by

D. R. Barr
G. K. Pooch
F. R. Richards

November 1978

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral T. F. Dedman
Superintendent

J. R. Borsting
Provost

This report is an outgrowth of work performed by the writers in designing experiments for evaluating specific command, control and communications technologies at the ACCAT facility located at NOSC in San Diego, California. The ACCAT Experimental Test and Evaluation Manager requested that we prepare a basic methodology manual, which would address experimentation issues, to help personnel in various disciplines understand the basic concepts and issues involved in the experimentation process.

This report represents the first volume of such an experimentation manual for ACCAT researchers and experimenters. The second volume discusses experimentation concepts in the context of a number of broad ACCAT technology areas.

This report was prepared by:

D. R. Barr, Professor
Department of Operations Research

G. K. Poock, Professor
Department of Operations Research

F. R. Richards, Associate Professor
Department of Operations Research

Reviewed by:

Michael G. Sovereign, Chairman
Department of Operations Research

Released by:

William M. Tolles
Dean of Research

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55-78-032	2. GOVT ACCESSION NO. 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) EXPERIMENTATION MANUAL. Part I. Experimentation Methodology.	5. TYPE OF REPORT & PERIOD COVERED Technical rept;
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) D. R. Barr, G. K. Brock, F. R. Richards	8. CONTRACT OR GRANT NUMBER(s) 52p.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62702E, N6600178WR00179
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940	12. REPORT DATE November 1978 13. NUMBER OF PAGES 47
14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

ACCAT,
Command and Control experimentation,
EXPERIMENTAL DESIGN

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Experimental Design Concepts applicable to ACCAT experimentation with C2 technologies are discussed. A number of considerations which should be included in the process of design, conduct and analyses of experiments are presented.

A 251 450

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

EXPERIMENTATION MANUAL

PART I: EXPERIMENTATION METHODOLOGY

by

D. R. Barr
G. K. Poock
F. R. Richards

ABSTRACT

Experimental Design Concepts applicable to ACCAT experimentation with C2 technologies are discussed. A number of considerations which should be included in the process of design, conduct and analyses of experiments are presented.

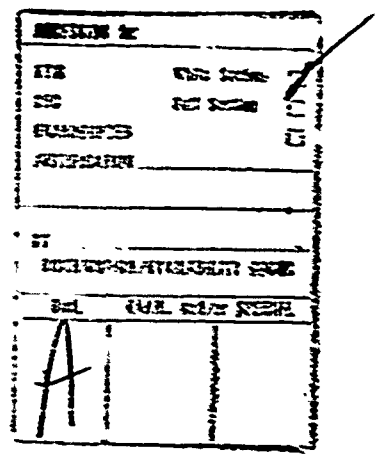

TABLE OF CONTENTS

EXPERIMENTATION MANUAL

PART I: EXPERIMENTATION METHODOLOGY

by

D. R. Barr
G. K. Poock
F. R. Richards

A. INTRODUCTION

The ACCAT project has two broad objectives: technology transfer and specification development. Achievement of these objectives, even over selected segments of the C2 arena, will require a wide spectrum of activities and equipments. In all cases, the goal is to obtain information useful to the C2 community, whose members range from the engineers and designers of hardware and software to the operational users of these assets. In order to be useful, this information, in the form of reports, recommendations and demonstrations from ACCAT, must be as accurate as possible within the experimentation resource constraints. Moreover, the information must be credible, and the quality (confidence level) of the results should be reasonably well established.

The breadth of this undertaking, in terms of the diverse technologies and equipments, large number of users with widely varying interests and responsibilities, diverse (and sometimes competing) goals and measures, implies that careful planning and

development of experiments in support of the ACCAT objectives is essential. The aforementioned breadth notwithstanding, there are many general features of the proposed experiments that are common to a large segment of the total experimentation effort. In what follows, we discuss several such features, first in very general terms, and later in terms of a number of broad technology areas.

This report is an outgrowth of work performed by the authors in connection with experimentation support to the ACCAT facility at NOSC in San Diego, California. Several related reports are listed at the end of this report (References 1 and 2). A glossary of statistical terms used in this report is included in Section H.

B. EXPERIMENTATION CONCEPTS

1. Levels of experimentation.

The degree of formalism in an experiment can fall anywhere on the continuum between very loose "free play" with only subjective judgment output to "highly structured" with complete specification of conduct of trials and output consisting of carefully measured system attributes. It is convenient to identify four broad bands of experimentation formalism; from least structured to most structured, they are:

a. Validation Experiments

"Experiments" consisting essentially of debugging software or hardware implementation of a technology. The output of such activities, also sometimes called "validation trials" is a determination of feasibility of the system to work to within fixed specifications. Frequently, another output that is desired is improvement or enhancement of the system under trial. An example of this type of experiment is exercising and debugging a computer program prior to placing it "on line" for operational testing or (later) use.

b. Demonstration Experiments

The term "demonstration" actually refers to the type of output ("report") from the experiment. Such experiments are frequently quite loosely structured, and output consists mostly of personal impressions in the minds of the users of the system under demonstration and outside observation of the trials. Such experiments may be somewhat more structured than the validation experiment, in that a scenario is followed or a set of activities is performed during the demonstration.

c. Assessment Experiments

Experiments in which trials are conducted over possibly a very wide range of conditions, with perhaps little control over sources of error. The goal is to gain an idea of how well the object of experimentation works, as judged in broad terms. Frequently, only subjective opinions of experimentors and observers are recorded as experimentation data.

d. Evaluation Experiments

This is the most rigorous type of experiment, with careful control of experimental conditions, with possibly a number of replications, usually followed by a formal analysis of numerical measurement data.

While many experiments will exhibit properties of several of these classifications, it is useful to adopt common terminology conveying, at least in general terms, the degree of formalism of each given experiment. It is probable that the more formalized type of experiments will be appropriate for answering narrower, more technically oriented questions; these might be described as "engineering questions." The broader, "looser" problems will probably be approached using the less formal types of experiments; many of these experiments could be described as concerning "operational questions."

2. Scope of Experiment

It is important to be selective about the factors, variables and conditions which are to be included in the experiment. This involves consideration of the resources available for the experiment, determination of the relevant and important aspects of the situation or technology to be investigated in the experiment, risks associated with the quality of inference

that will result from the experiment, and many other facets of the experiment and situation. If too broad a scope is selected (trying to cover too much territory), the result may be an inefficient experiment. For example, it is possible the desired inferences cannot be made with reasonable confidence, so the conclusions then have low credibility. The effects of factors, relationships among variables, etc., may go undetected because of the high error components in the observations (low signal to noise ratio). The other extreme, too narrow a scope, is equally inefficient. If a specific narrow detail is subjected to a very large amount of experimentation, the result may be a very definitive (and credible) conclusion which contributes little to general understanding or solves nothing of consequence. The population to which the inference applies may be so small that a definitive statement about the population may have little utility to the C2 community.

One approach which can be of value for the problem of determination of appropriate scope is to undertake relatively restricted experimentation at first, and to extend to more ambitious scope gradually, as a firm basis is established for the narrower scope. This constitutes "suboptimization," but may be feasible in situations where not enough is known to reliably determine appropriate scope directly. The experiment efficiency following this approach is traded away to gain insurance that one does not "bite off more than he can chew."

In any case, the scope of each experiment should be deliberately, and carefully, set. One should incorporate alternates in the experiment plan, when possible, so that scope can be downgraded or upgraded in response to results of early trials of the experiment.

3. Credibility of Results

It will do little good to run an experiment if no one believes the results. Whether we like it or not, an important part of an experiment is "selling the results." Of course, the best way to "sell" is to have a good product which meets a demand (we discuss these aspects elsewhere). But the consumer must also perceive the product is good and meets some of his needs. Thus, it is important not only to "discover the truth," but to do so in a way that others are compelled to the same conclusions. We mention this because there is a tendency for those familiar with a device or technology to arrive at subjective opinions about it, and to feel that there is little need for experimentation at a more formal level. We suggest that mere statements of opinions, possibly by individuals who have a "stake" in the project and who may be viewed by outside consumers of the experimentation results as being biased, may fail to be credible to these consumers. Such opinions or "expert judgments" may be quite accurate. An associated problem with such judgments

is the difficulty of assessing the quality (soundness, basis, confidence level) of such inferences.

Subjective judgments can certainly be of value, indeed in some situations they may be the only feasible measures to make. But self-proclamation by a researcher that he knows the answer may not be convincing to others if it is not backed up with more formal experimentation.

Generally, credibility of a conclusion depends not only on the quality of the procedure used to reach the conclusion, but also on how the case is presented and how the conclusion compares with preconception by the consumer. Generally, the highest credibility is attained using a formal level of experimentation with a definitive reporting technique, where the conclusion reached is "what the consumer expected (or wanted) to hear."

4. Determining What to Measure

Often the nature of the system under investigation determines, at least to a large extent, what can be measured during the experiment. Roughly speaking, we are concerned here with determination of what will be the dependent variables for the experiment. Usually, the more formalized types of experiments will involve mostly measurement of attributes through MOE's (measures of effectiveness), while the less formalized experiments will depend more heavily on MOP's (measurements of performance),

which may be numerically valued only in some limited cases. Often, the demonstration experiment will involve only subjective opinions, and these may not be even explicitly recorded. Rather, the demonstration may itself be the "report" of the experiment; conclusions and inferences may be confined to the minds of those who participate in the demonstration. Thus, while there are a number of general properties and qualities one would like his measures to have, the plain fact is the situation under study often involves constraints or difficulties which in effect determine the measures to be used. We measure in an experiment that which can be measured and that which seems reasonably related (perhaps through MOE's) to the goals of the experiment. Thus, in a C2 experiment we may measure things like times between certain events, error rates, flow rates, backlogs, etc., even though these measures may be only indirectly related to real operational problems and questions. We believe it makes sense to do this for various reasons we shall discuss below, including development of baseline information, calibration of the experimental apparatus and determining relationships between operational attributes of a system and design specifications for the system. It is often the case, therefore, that the formal measurements taken in an experiment seem only very remotely related to the real, significant, operational problems related to the experiment. But they should be made, because they are relatively inexpensive, and they do have values for reasons such as those mentioned above.

In addition, measures of a broader nature, including subjective judgments and assessment of utility should be undertaken and analyzed to the extent resources permit. We now pause to discuss some general aspects of MOE's and utilities.

5. MOE's and Utilities

One of the most important tasks of the experimentation planning phase is definition of data to be gathered. Objectives of experiments are frequently numerous, and to a varying extent, they compete for experimentation resources and interact with one another. If experimentation is to satisfy the objectives, measures of effectiveness (dependent variables) must be developed which can be reliably evaluated without prohibitive cost. These measures, either singly or in combination, provide measured results appropriate for use in satisfying the objectives.

The factors (independent variables) and their levels, such as technologies, scenario conditions and subject types, to be included in each trial must be selected so that the experiment objectives are met within the experimentation resources. This includes development of an experimental design specifying the combinations of factor levels under which trials will be made.

The data volume (sample size, trial time, test duration, etc.) must be controlled so that enough data is obtained to meet the test objectives, but wasteful "over-kill" is avoided.

Determination of sample size, for example, may depend upon assessing anticipated variability to be encountered in the various measures of effectiveness under the various combinations of factors included in the experimental design, together with assessing how such variablility affects satisfaction of the experiment objectives.

In order to meet the ACCAT objective of technology transfer for C2 systems, it is necessary ultimately to evaluate candidate systems in terms of their utilities in the Fleet. This requires judgments on the part of the "decision maker" (DM) concerning the anticipated pay-off (in his judgment) associated with using each system. There are several ways we might imagine this process taking place.

Idealistically, each candidate method or system (alternative) competing for selection for use in a given application area can be thought of as associating a consequence (outcome) with each scenario (state of nature). In turn, each consequence has a utility for the DM. Thus, one can view the alternative systems as being "prize functions" which associate with each possible state of nature a utility of the resulting consequence. In order to obtain his system utility for each alternative, the DM computes its expected utility, where expectation is taken with respect to the DM's subjective distribution of possible (future) states of nature. If more than one DM is involved in the decision concerning which of the alternatives to choose,

the individual system utilities for each DM may be smoothed or averaged. Possible approaches to the latter include use of the Delphi Technique, scoring and averaging, and appeal to some "expert" or "authoritative" DM for final judgment (Reference 3).

The role of experimentation and measurement of MOE's in this process is to give information about each system to each DM. The information in the form of measured values of MOE's (or a statistical summary based on such values) should be invariant with respect to the DM. Some MOE's are also invariant with respect to scenario, so the processes of obtaining expected utility and smoothing these over DM's is not necessary. For MOE's which do depend on the scenario, care must be taken to:

a) Use a broad range of scenarios in the experiment so as to provide the DM a basis for judging its contribution to, or impact on, expected system utility, and

b) document the scenarios used and report measured MOE values in the context of the scenarios under which they were obtained (see Reference 4 for one approach to this problem).

6. Determining and Reporting the Quality of Inferences

An important aspect of the more formal experiments is that, if well designed and executed, estimates can be made of the quality of the inferences being reported. This quality may be in any of several forms, including confidence levels, levels of significance and

operating characteristics of statistical tests, or variances or standard errors of estimates. Such information is a valuable part of the report of experimental results; it can add significantly to the credibility of conclusions based on the experimentation results. For the less formal experimentation types, it is usually not possible to make statistical statements regarding quality of inferences. In such cases, the credibility of inferences may be low.

7. Experimentation Plan

Questions concerning the nature of the experiment, its goals, experimentation units (subjects, equipments, etc.) to be used, resources required, and procedures to be used in conducting the experimentation trials should all be addressed in advance of performing trials. Experience shows that resources expended in experimentation planning usually more than pay for themselves in increased efficiency of the experiment.

For the more formalized types of experiments, the experimentation plan can be broken into three distinct parts: the experimental design, the schedule of trials, and the analysis plan. Clearly these parts are interrelated. For example, the analysis procedure depends critically upon the design and order of trials. The experimental design specifies the combinations of factors to be used for the various trials and the numbers of

trials to be performed within each such "cell" of the design matrix. The schedule of trials specifies the order in which trials required in the design matrix are actually conducted.

Development of a design which is efficient, and which leads to answers to relevant questions about the system under test, is usually a difficult task. It often requires efforts of a team of experiment planners, including statisticians, engineers familiar with testbed software and hardware, managers familiar with the test cycle, etc. Each experiment involves individual design problems which are not in common with other experiments. Thus, it is not reasonable to expect any "canned" design, no matter how successful in a past experiment, to be precisely relevant for use in another experiment. We therefore limit our discussion in this report to general principles of experimentation planning which experienced personnel consider in developing a design, schedule and analysis plan for a particular experiment. Applications of many of these principles to ACCAT evaluations are referred to in Part II of this report.

8. Experimentation Plan Documentation

The proposed design and plan for conducting each experiment, along with its objectives, requirement of resources, and anticipated results should be documented. This document should serve as a "blueprint" for conducting the experiment,

as well as serve as a source of information for individuals not directly involved in experiment planning. Such a document, suitably modified to reflect "evolution" in the plan, can serve as a foundation around which the final report of the experiment results can be written.

It is helpful to both readers and writers of the experimentation plan to use a common format for all plan documents. The following format is recommended.

Title of Experiment (or Series of Related Experiments)

EXECUTIVE SUMMARY

1. EXPERIMENT
   a. Title (specific)
   b. Number
2. OBJECTIVE
3. RESOURCES REQUIRED
4. GENERAL CONTEXT
   a. Concept and Need
   b. General Situation and Scenario
5. EVALUATION
   a. Data Collection (including what is to be measured)
   b. Anticipated Statistical Analysis of Data
   c. Anticipated Results
6. COMMENTS AND SPECIAL INSTRUCTIONS
7. APPENDICES (as appropriate)

C. EXPERIMENTAL DESIGN

1. Use a Design Plan and Established Design Principles

Making plans for efficient experiments requires proper use of information related to the systems under test. This includes such factors as the anticipated variability in the measures to be taken, and its impact upon design and sample size requirements. Advanced command and control system concepts frequently involve many objectives and attributes. Typically evaluations of various attributes are not optimally achieved by any single test plan. Thus it is often necessary to examine and analyze trade-offs among the "competing" objectives as they relate to experimentation procedures and measures, so that a reasonable compromise plan can be attained. The proposed conduct of experimentation must be carefully developed so that the procedure is both within the resources made available for experimentation and within the design goals. Avoidance of unnecessary confounding and attention to sound experimentation techniques so as to produce credible results are among the rewards to be gained with good test plans.

2. Use Common Features for Several Experiments

Frequently a series of experiments can be designed so as to involve use of similar techniques, concepts, procedures and tools. In such cases, it is obviously efficient to make

repeated use of these common features, since cost of development can be spread among the experiments. An example of this is development of experimentation technology for use in several experiments. This technology might involve development of timers to measure and record times elapsed between "milestone events," a questionnaire analysis procedure, and scenarios for use in conducting trials. With planning, experimentation technology can be developed so as to serve in more than one experiment.

3. Getting Data for Several Objectives in One Experiment

It is sometimes possible to get information about several aspects of a process or technology at the same time (within the same trial); where possible, this can be an efficient way to proceed. Some examples of this are:

a) Use "debugging" activities to get some formal measures on related variables, such as subject variation and process times.
b) Use subject "warm up" or indoctrination period to measure training and learning variables, and vice versa.
c) Assess data base reliability (or quality) while debugging the data base.
d) Use a procedure that measures the decision maker/operator interface variables, while also measuring operator/computer system variables.

4. Use a Sequential Approach

Often there is not a good basis available for selecting appropriate levels of factors prior to conducting trials in the experiment. In such cases, it may be useful to conduct testing sequentially, where experience and information from early trials can be used to help design later phases of the experiment. Care must be taken to plan this approach properly, so results from the various phases can be merged in the analysis phase. Frequently this approach consists of only two phases--a pilot phase and a "record trial" phase. Data from the pilot phase is used to calibrate the experimentation set-up, and to allow modifications in levels of factors and design, if necessary. Because of its importance, we discuss the idea of pilot trials in general terms in the following paragraph.

5. Use Pilot Trials

In many experiments, we wish to determine if there are differences in a subject's performance under changes in various levels of experimental factors such as color combination, target presentation method, data query method, etc. It is likely that the impact of changing one or more of these factors, in terms of what is measured, depends on the difficulty of the subject's processing task. Thus, our ability to discriminate between levels of the measured values depends on the difficulty of the information processing tasks. If the tasks are too easy, the

measures of effectiveness will be too high under all experimental conditions and no discrimination will be possible. On the other hand, if the tasks are too difficult, the measures of effectiveness will be very poor under all conditions, and again little information about the effects of the experimentation factors is gained.

A typical plot of discrimination (say, measurable difference in performance over changing levels of the factors, measured in signal-to-noise units) of a given MOE might be as shown in Figure 1 below, where processing content might be measured in bits per second, for example.

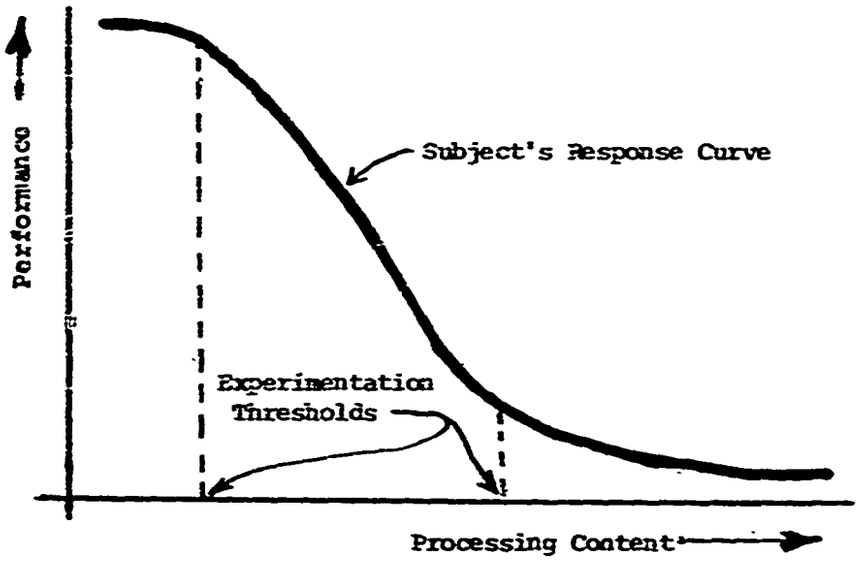

FIGURE 1. Best discrimination between effects of changing experimentation factors ("processing content") is achieved at midrange of processing difficulty where performance level (as measured by MOE's) is relatively sensitive to such changes. Experimentation outside the "thresholds" would require large expenditures of resources to establish the significance of differences in factors.

The experiment will yield relatively little information about effects of changing levels of the factors if the trials are conducted outside the threshold region shown above. For many experiments, we have at our disposal controls on processing content. We may not have a very good idea, however, at design time, of the levels of these controls that would provide trials within the threshold region. In such a case, we would recommend "pilot" trials be conducted (with typical subjects, if possible) in a trial and error mode to search for the threshold region. For example, in the record trials (as distinguished from the pilot trials) we might wish to hold fixed one of the processing content controls, say difficulty of the task. With a proper choice for this control, the other control, say "situation," is to be varied over levels which should cover the threshold region for some measured response variable (MOE). The pilot trials would allow appropriate choices for these controls. If, for example, the situation is controlled through selection of scenario "slides" to be displayed, it follows that the pilot trials should be undertaken as early as possible: in any case they must be completed before construction of the slides to be used in the record trials.

The concept of relating processing content to discrimination level through an operator's response curve has potential for use in various experiments. It is of interest for the C2 application to determine how wide the threshold of

discrimination is, in terms of the range of command control situations actually encountered in practice. For example a very narrow threshold, which falls well off the level encountered in a crisis situation, might imply "it doesn't matter" what level one uses. The variation in shape and location of response curves between subjects is also of interest. For example, very large differences in threshold location for different subjects might suggest there is not high operational value in attempting to select a "best" control combination.

6. Set Experimentation Priorities

It is important to allocate resources in an experiment at levels that are commensurate with the importance of the information anticipated to come out of the experiment. Thus, a high valued objective for which there is reasonable expectation of getting useful results through experimentation might be allocated priority over an experiment for which the objective is not of great importance, or for which there is high risk that useful information will not be provided by the experiment.

Experimentation prioritization depends not only upon experiment objectives and risks but also upon experimental design and analysis requirements. In many experimental designs, the efficiency of the experiment and the adequacy of analysis of resulting data may be seriously affected by omission of

certain trials, but much less seriously affected by omission of others. For example, in a factorial design with analysis of variance, having certain cells empty may render important parameters nonestimable. In such cases it might be possible to forego replication in certain other cells in order to devote test time and effort to obtaining at least one observation in the critical cells.

If, due to testbed availability restrictions, for example, it becomes necessary to discontinue experimentation ahead of plan, clearly any remaining trials should be carried out so as to first provide answers to the most important objectives. On the other hand, it may become evident, as experimentation proceeds, that conclusions with the required degree of confidence can be made before all of the planned trials are completed. In such cases, where testbed time can be used more effectively in other ways, experimentation plans should call for alternative use of experimentation effort.

Certain general principles of experimental design apply to the problem of experimentation prioritization, including:

a) Testing at extreme conditions first, in order to determine whether trials under intermediate conditions are likely to be informative.

b) Conducting "base line" trials with known systems or subsystems might be de-emphasized (but not eliminated) if necessary. In some cases, credibility of the experiment results requires

some base line trials to validate the experimentation set-up and to certify its calibration for comparisons with other sources of data.

c) Sequential testing is known to minimize expected sample sizes in some cases. This could provide a basis for prioritization.

Meeting scheduling requirements and responding to changing availability conditions should create, but suggest some flexibility in, planned experiment priorities.

7. Determine Experiment Resource Availability

The impact of experiment resource nonavailability can be reduced. Obviously, availability of technologies, equipments, subjects and testbed facilities is a major constraint that must be satisfied in conducting experiments. Thus, obtaining estimates of periods of availability, both directly for experimentation purposes, and remotely through conducting limited nonscheduled trials concurrently with other primary use of the resources, is of great importance.

Experimentation may not always proceed as planned (due to conflicts, personnel changes, equipment failure, etc.), however, careful scheduling of experimentation resources, together with planned alternatives with perhaps reduced configurations may make it possible to obtain useful data, even in "down-times."

It is essential that realistic est ates of time requirements be made. If these estimates are inconsistent with anticipated availability of resources, the experiment objectives should be revised. In some cases, if such revision is not reasonable, a "no-go" decision may be appropriate.

In some cases where experiment resource availability and test time requirements cannot be adequately determined for planning purposes, it may be possible to perform a modest preliminary program with one aim being assessment of these quantities.

8. Define the Experimentation Level, Scope and Duration of Experiment

The scope and duration required is a function of experiment objectives and constraints on resources. Experiment planning includes determining an initial set of objectives and searching for ways to meet these objectives. Frequently, the initial set proves too optimistic in that they cannot be met with the resources available. Consequently the process may involve re-evaluation of objectives to a more feasible (modest) set, with search for the best ways of accomplishment. The scope and duration of experimentation is often a result of compromise between excessive objectives and insufficient resources to meet them.

The scope, duration and level of experimentation must be of sufficient size to include a range of conditions and sample sizes adequate to (1) provide reasonable statistical confidence level, (2) give the experiment results credibility, (3) give a useful range of applicability of results, and (4) provide a basis for prediction of system performance and features over conditions other than those specifically included in the experiment.

9. Coordinate Testbed and Scientific Community Efforts in Planning and Conducting Experiments

System and subsystem designers and operators and members of the scientific community can provide information valuable in the development of experiment plans, as well as assist through involvement in experiment implementation, data collection, and interpretation of analysis results. Members of the scientific community associated with the system or subsystems under investigation may possess information or knowledge of the systems useful in planning the experiments. For example, they may possess information about system reliability and appropriate measures of effectiveness and their variability. In addition such individuals may be involved, either directly or indirectly, with the systems while they are undergoing experimentation. For such reasons, close coordination among

these communities should be maintained throughout the periods of planning, conduct, analysis and reporting.

System operators and other personnel associated with the systems under test may have accumulated valuable experience with certain aspects of the systems. In addition, it is possible some of these people will be present during experimentation, possibly even as experimentation subjects or operators in some cases. Again, the advantages of close coordination are evident. Trials may, at least in part, take place in conjunction with other system activities in which members of the scientific community may be involved. For example, some system testing may be planned as only part of overall exercises and evaluation in which contractor and scientific personnel must be on board.

Finally, since systems under examination may be state-of-the-art, knowledge of selected experts is valuable in planning and conducting tests.

10. Make On-Site Trial/Exercise Evaluations

An important step in any experiment procedure is to conduct on-site evaluations of the procedure and perform preliminary data reduction and analysis. On-site evaluation of experimentation, as it is conducted, performed by experienced test personnel, is necessary to validate trials and to verify conditions of the experiment design are being met.

An important result of "real time" evaluation of experimentation, on an on-going basis, is the verification that the procedure is in control--that is, that instrumentation is up, players are properly oriented, subsystems are operating, factors are at proper levels, etc. On-site evaluation is needed for making decisions about conducting trials with systems in reduced configuration, making changes in the experimental design, or changing prioritization.

Information obtained from on-site trial/exercise evaluation, based on in-depth discussions with operator personnel, test personnel and members of the scientific community, may be of critical value in assessing results of the experiment. Such information may form an important part of the data collected.

11. Compare Test Performance with Anticipated Values

The extent and manner in which actual experiment data differs from anticipated or predicted values impacts upon the experiment's credibility and adequacy, and it impacts on design of subsequent experiments. In cases where observed experimentation data are not significantly different from predicted values, it may be desirable to reduce sample sizes in certain cells of the design and to allocate experimentation effort over a wider range of conditions, depending upon the basis of the predictions.

In such situations, the credibility of experiment results may be relatively easy to establish and extrapolation to test conditions other than those originally planned may be possible.

In cases of substantial disagreement between observed experiment values and predicted values, more intense experimentation in a few cells may be necessary to reduce variance to a point where little doubt of the significance of the difference remains. Investigation of the possible sources of the disagreement is desirable; otherwise, credibility of the experiment results may be in jeopardy, and extrapolation of results may be untenable.

Design of subsequent experiments should take into account experiences in previous related trials. Artificialities in the experiment procedure and in the system under investigation should be assessed and compensated for, or at least accounted for.

12. Schedule Trials

"Proper" scheduling of trials is important to reduce effects of unknown and possibly unforeseen factors that might otherwise become confounded with the experimentation factors of interest. For example, one frequently encounters a "time effect" in experiments wherein effects such as motivation of personnel, learning by operators, long-term seasonal effects (weather,

visibility, day length, etc.) and gradual upgrading of systems involved, may influence effectiveness measures in the experiment.

If there is no difficulty in doing so, a generally accepted method of scheduling trials is to select trials in random order. Obviously, this may be prohibitively expensive or otherwise infeasible in some experiments. Thus we may be forced to compromise away from the "completely randomized" schedule. In such cases, it seems reasonable to conduct as many trials as possible with each given experimentation set-up. However, within a set up, it is desirable to select the order of the trials at random, to the extent possible. Thus, for example, within a technology type, the design might call for a dozen trials, one for each combination of two scenarios × three subjects × two display devices. We might plan to run these 12 trials within one period in which the technology is "up." However, the order of execution might be selected using a table of random numbers. The 12 trials would be conducted in the order in which their codes were drawn from the random number table. For example, if it is not feasible to change scenarios frequently during a technology up period, plan to run all of the trials of one scenario first. Then randomize the order of subjects and displays (6 trials) for each of the scenarios, as discussed above.

The randomization should be carried out formally for each cell (technology or technology/scenario combination in the preceding example) in each experiment. Once a design has been

selected, the trial schedule can be developed in advance of actual experimentation. If extra trials become available with a technology, they should be run, replicating some earlier trial conditions or filling in combinations not previously run.

D. EXPERIMENTATION GOALS

We have discussed the importance of setting experimentation goals as part of the planning sequence. In what follows, we give an outline of several types of goals that may be involved in experiments.

1. Comparison

We may wish to determine which of the several alternatives is "best" in some sense.

2. Establishing a Base Line

Experimentation may be undertaken to document the attributes of some standard technology, or of a technology currently in use. This will provide a standard context in which to judge related new technologies, as well as to allow calibration of new experimentation set-ups.

3. Establishing Feasibility

It may be desired merely to demonstrate that a concept or projected procedure or newly developed technology can work.

4. Determining Specifications

One difficulty in conducting experiments with emerging technology is that there are no specifications to "test against." As mentioned above, in addition to its technology transfer objectives, ACCAT has the objective of establishing specifications for use in continuing development of the technologies under investigation. In order to determine specifications for a system, it is desirable to relate measured attributes of the system to operational characteristics of the system. This implies that measurements must be made of technical quantities, at an engineering level, concurrently with observations of simulated (and, where possible, actual) operational characteristics of the systems under investigation. In many cases, measurement of the technical quantities will require that "probe points" are designed into the system when it is developed. For example, measurement of flow rates and cycle times within a software package requires that provision for counters and timers be present in the software, or at least that such devices can be hooked to the software at desired points. It follows that success in specification determination requires that technology developers/contractors make

measuring devices, or accommodations for them, available as part of the technology package.

5. Demonstrations

Demonstrations may be considered to be a combination of low level experiment and report of results. The object may range from demonstrating feasibility of a concept to conveying information to those involved in the demonstration.

6. Evaluation of Effects of Factors

It may be desired to know whether certain factors have a significant effect on measures of effectiveness, and if so, how the effects are characterized.

7. Removing Effects of Factors

Some factors are not of interest in themselves, but have potential for influencing measurements in an experiment. Such factors are sometimes referred to as "nuisance effects." It is desirable to measure values of nuisance effects so that in the data analysis process the effects can be accounted for and removed from the effects of interest. Failure to make measurements on nuisance variables can mean their effects add to the apparent noise level in the experiment, making significance of

factors of interest in the experiment needlessly low. For example, we might measure typing ability of a subject in an experiment involving query systems, not because we are interested in typing effects per se, but because we believe different subjects may have different typing skills, and this may affect the values measured on the variables of interest.

8. Assessment of Utility

In our discussion of MOE's and their role in the decision making task, we referred to utilities of consequences which, when averaged over scenarios and smoothed over decision makers, gave utilities of technologies. It may be possible to numerically evaluate utilities of some of the technologies under investigation, following this scheme. There has been a considerable amount of research on this subject reported in the technical journals in recent years. It appears worthwhile to devote some resources to examining the feasibility of utility assessment in the ACCAT context.

E. HANDLING DATA

It is important to plan data handling in advance of experimentation. The nature of the data, the volume of data to be obtained and anticipated analysis approaches are facets of data handling that must be considered. Questions concerning

use of computers in data reduction and analysis impact on the form in which data are to be stored. Failure to store data in suitable form may render it practically useless, even though considerable expense and manpower were expended in gathering "good" data. For example, data stored in hard copy form is extremely difficult to convert to machine readable form, even if there is only a modest volume of data. For large volumes, it becomes infeasible to convert hard copy to machine readable form. The old adage "an ounce of prevention is worth a pound of cure" is certainly appropriate for this facet of experimentation planning.

Selection of procedures for data gathering resulting from experimentation depends on many factors, including the anticipated nature of the data, the volume of the data, operational considerations, soundness from an experimental design point of view, reliability and cost.

The nature and volume of data to be obtained in planned tests can vary greatly. For example, simulated tracking of an aircraft may generate large quantities of instrumentation data measured and recorded "automatically," whereas determining an operators' opinion about a piece of equipment may involve questionnaire data in relatively small amounts in some cases recorded "by hand." Security and safety considerations may impact upon data collection and test conduct. For example, safety considerations might require development of data from

simulated laser application rather than actual use of a laser in the experiment.

Reliability and cost of experimentation system hardware and software may impact upon data collection and trial conduct. For example, if a laser tracking device planned for use in the test proves to be unreliable, it may become necessary to use different hardware or even a different test concept.

In order to properly plan to accommodate and handle data produced in an experiment, a number of aspects must be kept in mind, including:

i) automation in recording, handling and analysis
ii) storage method
iii) storage format and identification
iv) anticipated analysis procedure
v) schedule of experiment, including planning, execution, analysis and reporting phases.

These aspects should be addressed and documented in the experimentation plan.

F. ANALYSIS OF RESULTS

1. Analysis

Proper analysis of the experimentation results is necessary to meet the experiment objectives as well as gain insight into the behavior of the systems under examination (for example,

examination of interactions among factors included in the trials). Selection of analysis procedures appropriate for use with test results depends not only upon the planned experimental design, but also upon possibly unforeseen aspects of the trials and the data themselves. Occasionally unique analysis techniques must be developed to meet the experiment objective with the data obtained.

While the analysis plan should be an integral part of experimental design, analyses actually performed depend also upon the nature of the data obtained, and the circumstances under which they are collected. An analysis plan should be developed along with other aspects of the experimental design. The experimental design should be selected with proposed analysis procedures, as well as experiment objectives and resource constraints in mind. Frequently a major portion of the planned analyses are parametric analyses of data obtained on the measures of effectiveness. For example, multiple regression and analysis of variance may be planned for use with subject performance-time data. As a result of unforeseen difficulties in following the experimentation plan, or unanticipated responses or behavior in the measures taken, the planned analyses may be inappropriate for the data actually obtained. For example, limitation of times allowed for completion of a certain task produces truncated data for which standard parametric procedures may be inappropriate.

Assessment of the tenability of assumptions required for proposed analyses is an essential part of the analysis procedure. Special analysis procedures different from, or in addition to, those planned may be required due to the aforementioned circumstances. For example, nonparametric analyses may be required if the parametric assumptions appear incompatible with the data obtained.

Development of special analysis procedures may require research of the statistical literature, generation of new methodology, or adaptation of standard analysis procedures to the given experiment situation. For example, a transformation of observed "time to completion" data may be required to stabilize variance so analysis of variance may be applied.

It is important to plan and conduct analyses at a level appropriate for accomplishing the objectives of the experiment (within resource constraints, as always). It makes little sense to perform a lengthy, involved, time consuming analysis with data that are "poor" or were obtained with a low-priority (passing interest) experiment. In such situations, a "quick and dirty" analysis, perhaps consisting only of development of summary statistics, may suffice. By the same token, it is not efficient to allocate resources for only a low level analysis with an experiment of high interest for which good data were obtained with extensive experimentation

effort. Planners of experiments should be aware that good analysis takes time and effort, and to shortcut this process may unnecessarily degrade the experimentation results.

2. Documentation.

Documentation of experimentation results, in the form of reports, records and data sets, should be provided at various stages of the experimentation procedure.

An important part of the experimentation plan is the establishment of documentation requirements. Normally, documentation for each phase of an experimentation sequence is provided upon completion of the phase. Time requirements for each report should be planned well in advance.

Determination of who is responsible for each report, or portion thereof, with appropriate monitoring and control, is important. Frequently, cooperation of experimentation personnel, analysts, members of the scientific community and contractors is required for preparation of accurate, timely documentation and reports.

A standard format for the final report is desirable. As mentioned previously, the experimentation plan document can serve as the framework around which the final report can be written.

G. CONCLUSIONS

Good experimentation requires careful planning by personnel familiar with the systems under investigation and the principles of experimental design. There are many, sometimes competing, experimental design goals. There is a large number of considerations which must be dealt with in the design plan. While there may be various general aspects of different experiments which are held in common, there are virtually always signficant differences as well. It is thus unreasonable to expect a "cook-book" approach to experimentation to be successful. Rather, it is efficient to expend a portion of the total resources available on development of an experimentation plan tailored to each specific experiment. Similarly, the analysis of results from the experiment will usually require special (i.e., nonroutine) effort, for which resources should be programmed.

In this report, we have discussed a number of considerations which should be included in the process of design, conduct and analysis of experiments.

H. GLOSSARY OF STATISTICAL TERMS USED IN THIS REPORT

(Note: more complete glossaries are given in References 5 and 6.)

1. Analysis of Variance--A statistical analysis procedure used to test whether changing values of certain input parameters or conditions have significant effect on the output mean value.

2. Cell--A specific combination of input parameter values and experimentation conditions, under which one or more output values are to be observed.

3. Confidence Level--A statistical tolerance value indicating the rate of making true conclusions concerning unknown parameters.

4. Confounding--A situation in which the individual effects of two or more potential sources of change in output values cannot be distinguished.

5. Discrimination--The process of determining separate individual sources of observed effects or of separating potential sources of such effects into groups which share some common characteristics.

6. Estimable--The ability to determine, through analyses of the experimentation data, the contribution of a source or parameter to the output mean.

7. Factorial Design--An experimentation arrangement in which several parameters or sources of effects are varied, such that each source level or parameter value occurs with all combinations of the other levels or values.

8. Factors--Parameters or possible sources of effects on the output variables.

9. Inference--The process of drawing conclusions about a system, often based on analysis of data obtained from the system under specified conditions.

10. Level of Significance--The type one error rate (α) in a statistical test of hypotheses at which the observed test value would just lead to rejection.

11. Multiple Regression--A statistical procedure of fitting a linear function of several independent variables to observed dependent variable data.

12. Nonparametric Analysis--Statistical analysis in which only weak assumptions about the theoretical distribution of the dependent variables are made.

13. Parametric Analysis--Statistical analysis in which a specific distribution form is assumed to hold.

14. Replication--A subsequent value or set of values of the dependent variable obtained under the same conditions as previous values or sets of values.

15. Trials--Portions of an experiment.

16. Variance--A measure of the dispersion of a statistical distribution.

I. REFERENCES

1. Barr, D.R., G. K. Poock and F. R. Richards," Experimental Design and Analyses for Initial ACCAT Test Bed Experimental Demonstrations," Naval Postgraduate Technical Report NPS55-77-21 (1977).

2. Miller, H. G., D. R. Barr, G. K. Poock and F. R. Richards, "Experimentation, Research, and Operational Concepts for the Naval Postgraduate School Remote Site Module," Naval Ocean Systems Center Report (15 April, 1978).

3. Pill, J., "The Delphi Method...," Socio-Economic Planning Sciences, Vol. 5, 57-71 (1971).

4. Keeney, Ralph L. and Howard Raiffa, "Decisions with Multiple Objectives: Preferences and Value Tradeoffs," John Wiley and Sons, New York (1976).

5. Kirk, Roger E., "Experimental Design: Procedures for the Behavioral Sciences," Brooks/Cole Publishing Company, Belmont (1968).

6. Hicks, Charles R., "Fundamental Concepts in the Design of Experiments," Holt, Rinehart and Winston, New York (1973).

DISTRIBUTION LIST

	No. of Copies
Defense Documentation Center Cameron Station Alexandria, VA 22314	2
Library, Code 0212 Naval Postgraduate School Monterey, CA 93940	2
Library, Code 55 Naval Postgraduate School Monterey, CA 93940	1
Office of Research Administration, Code 012A Naval Postgraduate School Monterey, CA 93940	1
Defense Logistics Studies Information Exchange (DLSIE) Fort Lee, VA 23801	1
Dean of Research Code 012 Naval Postgraduate School Monterey, CA 93940	1
Dr. K. T. Wallenius Department of Mathematical Sciences Clemson University Clemson, S.C. 29631	1
Dr. Tom Varley Office of Naval Research Arlington, VA 22217	1
CDR Ron James Office of Naval Research Arlington, VA 22217	1
Deputy Commander, Operational Test and Evaluation Force, Pacific Naval Air Station, North Island San Diego, CA 92135	1

DISTRIBUTION LIST

	No. of Copies
Mr. Bill Adams DEPCOMOPTEVFORPAC NAS, North Island San Diego, CA 92135	1
Commander, Operational Test and Evaluation Force Naval Station Norfolk, VA 23511	1
Mr. Jim Duff, Code 223 COMOPTEVFOR Naval Station Norfolk, VA 23511	1
Don Belford NWTC VX 5 China Lake, CA 93501	1
Mr. R. T. Rains Advanced Systems Test and Analysis Department Naval Ship Weapon Systems Engineering Station Port Hueneme, CA 93043	1
Dr. Bartels Naval Underwater Systems Center New London, CT 06340	1
Mr. Glen E. Hornbaker Naval Surface Weapons Center Armaments Development Dept. Dahlgren, VA 22448	1
Mr. Carl Hyndon Naval Surface Weapons Center Operations Analysis Group Dahlgren, VA 22448	1
Mr. Charles M. Merrow Naval Ocean Systems Center Operations Analysis Group San Diego, CA 93152	1
Mr. Marshall John Tino Naval Surface Weapon Center Ordnance Systems Assessment Division Silver Springs, MD 20910	1

DISTRIBUTION LIST

	No. of Copies
Mr. Harry Norton Pacific Missile Test Center Code 1020 Pt. Mugu, CA 93042	1
Dr. S. Z. Mikhail Naval Oceans Systems Center Operations Analysis, Code 125 San Diego, CA 92152	1
Ken Moll Strategy Co. 1901 North Fort Myer Dr., Suite 809 Arlington, VA 22209	1
Ted Bean Science Applications, Inc. 8400 Westpark Drive McLean, VA 22102	1
Robert Wright MAD, PMTC Point Mugu, CA 93042	1
William F. Trisler Code 05431 NAVELEX Washington, D.C. 20360	1
Dr. H. W. Sinaiko 801 Pitt St. Alexandria, VA 22314	1
Hal Miller Deparment of the Navy NAVELEX PME108-3 Washington, D.C. 20360	3
Bill Carper NOSC, Code 832 San Diego, CA 92152	5

DISTRIBUTION LIST

	No. of Copies
Naval Postgraduate School Monterey, CA. 93940 Attn:	
D. R. Barr, Code 55Bn	25
F. R. Richards, Code 55Rh	10
G. K. Poock, Code 55Pk	5
R. J. Stampfel, Code 55	1
J. M. Wozencraft, Code 74	1
R. J. Roland, Code 52Rl	1